WP 30
24 February 2010

ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE
EUROPEAN UNION (EUROSTAT)

ORGANISATION FOR ECONOMIC COOPERATION
AND DEVELOPMENT (OECD)
STATISTICS DIRECTORATE

Work Session on Statistical Metadata (METIS)
(Geneva, Switzerland, 10-12 March 2010)

METIS: 20 YEARS OF PROGRESS

Paper prepared by the METIS Steering Group

Invited Paper

Submitted by US Bureau of Labor Statistics[1]

I. INTRODUCTION

1. This year marks the 20th anniversary of the first meeting of the METIS working group. As a result, the METIS Steering Group thought it was a good opportunity to review the work of the group over that time and mark its progress. This paper contains a view of the progress that has been made.

2. The basis for the international work on Statistical Metadata was created under the framework of a UN Development Program (UNDP) Regional Statistical Computing Project (SCP) between 1983 and 1989. The project was financially supported by the UNDP, and the UNECE Statistical Division was responsible for the project management. Originally, 19 countries participated in the project (Bulgaria, Canada, former Czechoslovakia, Denmark, Finland, France, former German Democratic Republic, Greece, Hungary, Ireland, Italy, Netherlands, Poland, Romania, Spain, Sweden, Turkey, United Kingdom, and former Yugoslavia).

3. The outcome of the initial work was that 4 SCP Joint Groups were established. They covered the following topics:

- Statistical Table Generation,
- Statistical Data Editing,
- Statistical Metainformation System (METIS),
- Statistical Data Base Management.

4. Since 1990, the work program and international cooperation on METIS and other SCP topics have been integrated into the program of work of the Conference of European Statisticians (CES). The original framework focused on NSOs, but now under the CES, the focus expanded to include ISOs, such as ILO, FAO, IMF, WTO, Eurostat, and OECD. In addition, early on, the NSO participation was limited to ECE countries;

[1] Prepared by Daniel Gillman – The opinions expressed in this paper are due to the author and do not necessarily represent official policies of the US Bureau of Labor Statistics

however that has now been expanded to include any UN country, international statistical organization, and central banks.

5. As stated above, the UNECE was originally given the responsibility for coordinating the work of METIS. As of the 2002 meeting, Eurostat was added as a sponsor, and OECD was added as another sponsor for the 2004 meeting. This arrangement continues.

6. In its 20 years, METIS has had a rich history and progression of ideas. Rather than going through each meeting and discussing papers and results, the paper contains a look at progress through changes based on several themes. The themes chosen are meetings and participation, planning, technology, terminology, formalism, and implementation. Each of these is described as follows:

- Meetings and Participation – This looks at basic meeting and participation characteristics, and how these have changed over time
- Planning – We trace the changes in how meetings are planned and organized
- Technology – The evolution of METIS work and the outlook on metadata management systems as affected by changes in technology
- Terminology – The ever increasing importance of terminology management is described along with its evolution in METIS
- Formalism – Attempts to formally describe metadata have been promoted over the years, and we sketch the main threads within METIS
- Implementation – The focus on actual system development and how that affects the perception of METIS

7. The paper contains discussion on each of the 6 themes, in the order given. The selection of the themes is due entirely to the author.

II. MEETINGS AND PARTICIPATION

8. This latest meeting (March 2010) marks the 15th time the group has met, either in a full meeting or in an interim capacity. At first, full meetings were held every year from 1991 until 1994; between 1994 and 1996 there was a 2 year gap; they were held about every year and a half between 1996 and 2002; and since then the full meetings are held every 2 years. Beginning in 2007, an interim meeting was held between the biennial full meetings. The interim meetings are focused on one or two topics. See Table 1.

9. It is clear from Table 1 that the trend in participation over time is on the rise. One anomaly is the 2000 meeting in Washington, DC, but travel to the US was obviously an impediment for many European and other countries. All other meetings have occurred in Europe, so travel is much easier for most participants. Ironically, the US meeting was very well attended in terms of the number of people because the full US statistical community had easy access to the meeting.

10. As more organizations and people attend the meetings, the meetings themselves have become more formal. This is a natural reaction to the needs of managing a larger group; however some of the intimacy of the smaller groups has been lost. This may not matter so much, as the character of the work itself has changed over time, too, as we will discuss.

11. Another area where the meetings and participants have been greatly affected is with the Internet. As late as the 1994 meeting, not all statistical offices had a Web site nor did all employees have email addresses. The 1998 meeting was prepared and conducted using a Web site for the first time. All the papers were electronically available, including information and meeting notices that were difficult to transmit in the past.

12. As Internet technology continues to evolve, the use of new technology is adopted. The Task Force for the Common Metadata Framework (CMF)[2] uses a Wiki to develop their documents. Wikis and other

[2] http://www1.unece.org/stat/platform/display/metis/The+Common+Metadata+Framework

collaboration Web sites are in increasingly common use throughout the METIS community. Email is used constantly as a means to announce meetings and share documents among the group. Web sites are so commonly used, that everyone expects that documents will be kept there. In a short time, the Internet has changed the work of METIS considerably.

Year	Month	City	Type	NSOs	ISOs	Observer
1991	Nov	Geneva	F	[3]	-	-
1992	Sept	Geneva	F	[4]	-	-
1993	Nov	Geneva	F	17	4	1
1994	Nov	Geneva	F	19	4	1
1996	Oct	Berlin	F	24 [5]	6	3
1998	Feb	Geneva	F	20	10	2
1999	Sept	Geneva	F	27	8	1
2000	Nov	Washington	F	18	10	0
2002	Mar	Luxembourg	F	28	6	2
2004	Feb	Geneva	F	28	11	1
2006	Apr	Geneva	F	38	12	2
2007	Jul	Vienna	I	23	5	1
2008	Apr	Luxembourg	F	33 [6]	8	1
2009	Mar	Lisbon	I	31 [7]	4	0
2010	Mar	Geneva	F	39 [8]	13	3

Table 1 - Meeting Dates and Participation

III. PLANNING

13. At each METIS meeting, a group would be formed during the session to consider future work. This group would formulate ideas based on the content and discussion in the current meeting and present them to the entire work session for approval, and these ideas would be incorporated into the agenda as topics for the next meeting. This format was used until the 2004 meeting, and this change will be discussed further, below.

14. During the 1998 meeting, the participants recommended that a Steering Group be formed that would work closely with the secretariat to determine the make-up for subsequent meetings. This changed the planning for each session dramatically. The meetings were growing in size, and given the fixed 3 day meeting schedule, it was clear there soon would not be enough time to hear everyone speak. The Steering Group members felt they could maximize people's time at the meeting by inviting some to talk on important topics. Other contributed papers would still be accepted, but the amount of time allotted to each of those speakers would be limited. This way the important ideas would be presented, and everyone else would have some time to present their work, too.

15. The Steering Group would take the recommendations for future work from the previous meeting and determine several relevant speakers to give invited papers on each topic. Then, any interested statistical office was free to contribute papers to a topic of its choice. The proviso was the presenters of the contributed papers would be given less time to speak. It appears this arrangement has worked fairly well despite the fact that the time for some speakers to make their presentations is reduced.

16. A change occurred during the 2004 meeting, where the group tasked to determine future directions recommended the development of a common framework for metadata. This arose from reports given during

[3] Not enough information is available to make estimates
[4] Not enough information is available to make estimates
[5] The United States additionally participated by telephone conference for part of the meeting.
[6] Australia participated via video conference.
[7] Numbers estimated from registered participant list
[8] Numbers estimated from registered participant list

that meeting on the results of the MetaNet project. During the next year, the Common Metadata Framework (CMF) was advanced, and the subsequent meetings (2006, 2008, and 2010) were all organized around the sections of the CMF.

17. With the advent of the CMF, the work of METIS is tightly focused on 4 major areas. This means planning for future direction at the METIS meetings is greatly reduced, and CES has goals by which to measure the progress of the work. Additionally, the work of the interim meetings is focused on one or two of the sections of the CMF.

IV. TECHNOLOGY

18. Computer technology changes at a very rapid pace, and there have been many changes since the first meetings of METIS. As noted above, the advent of the use of the Internet happened after the first METIS meetings were convened, and it has had a significant impact on the operations of the meetings since. Of course, many of the advances involving the Internet also include huge increases in raw computing power and storage. Here, we will look at some of the major changes on both METIS and the work on metadata management systems.

A. **Effects on METIS**

19. The Internet has had a huge impact on the way METIS meetings are planned and conducted. Prior to the 1996 meeting, not all participants had email, papers and notices were transmitted through FAX or by regular mail, not all statistical offices had Web sites, and the process of planning a meeting was much slower. Participants could expect hard copies of all papers and supporting documents to be provided in the meeting room.

20. By 1998, the UNECE could manage the papers for each meeting on the Web, and a Web site was established for each meeting. The meeting pages are linked to http://www.unece.org/stats. Papers were made available via the Web site rather than being printed. Planning for the meeting could now be conducted via email, and this made the Steering Group a feasible option.

21. As the Web evolved further, and Web 2.0 was established, Web sites supporting collaboration became available. This new functionality allowed the work of the Task Force on the CMF to move forward. A Wiki was established for consolidating this work. The collaborative nature of the Wiki software enables much faster sharing and development of the work of the Task Force.

B. **Effects on Metadata Systems**

22. The Internet has also dramatically changed the work of statistical offices. The changes have encouraged and forced the development of more sophisticated metadata systems to support these changes. In addition to changes to the Internet, raw computing power – processing speed, storage capacity, and software sophistication – has increase dramatically as well. The computing capacity of the laptop used to create this paper far exceeds that of the Sperry Univac machine the US Census Bureau used to process the entire 1950 US Census. So, the changes over the years have been dramatic, and here we will explore some of these and their consequences to metadata systems.

23. In 1990, data sets were still available on large tapes, but increasingly they were put onto CD-ROM disks, and smaller data sets often copied onto floppy disks. The huge (at that time) size of a CD-ROM changed the way people thought about disseminating data. These disks were small, and they could be inserted into a small mail pouch and sent to anyone who requested one. By comparison, the old tapes were large, heavy, and special equipment was needed to read them. The ubiquitous PCs had drives built into them to read CD-ROM disks easily.

24. The much larger storage capacity of CD-ROMs meant that it was now possible to include software to read and select portions of the data contained on the drive. However, this also meant that a system for describing the data was also needed, and metadata systems for describing data on the CD-ROM were designed. This more narrow focus was a function of the technology, yet it did not coincide much of the time with the theory of metadata systems already in existence.

25. Even in the early METIS meetings, it is clear the theory of how metadata ought to be managed and used supported the desirability of describing all the surveys and data produced in a statistical office. The papers by Bo Sundgren, Josef Olenski, and others attest to this. However, the technology at this time mostly focused developers on the problems of describing a few data sets at once. This was changed little from the time when data were mostly on tapes.

26. Change at this time was rapid, and the CD-ROM solution to data dissemination had a short time in the limelight. By 1994, most statistical offices were connected to the Internet, and the use of the Web followed right behind Internet connection. These changes brought a new focus to metadata systems and data dissemination. Now, much of the data the office produced could be made available on a Web site through protocols such as FTP and HTTP. Still, the largest data sets were not suitable, as they were too large to download given the existing connection speeds.

27. Now the focus began to switch towards the broader problem of building a “corporate” metadata system for all the surveys an office manages. As offices could make ever larger data sets available for download, the needs of metadata systems increased. However, a desire to build much larger systems met with new challenges, and in 1999 the idea that building “cathedrals” should not be an initial goal – introduced by Jostein Ryssevik – was an important outcome. The notion of building a system up step by step is an old strategy, but it required repeating.

28. Over the last 10 years, connection speeds have increased to the point where it is feasible to download almost any data set a statistical office offers. Many offices also offer good metadata to go along with them. Now, the challenges are around services, software designed to work on the Web to find, manipulate, and integrate data from different sources. The change is toward more standardization and building automatic means for finding and manipulating data, therefore the newest demands on metadata systems are to make the metadata “active”– able to be processed by the computer – and using a standard model rather than to keep metadata “passive” – for human use – and using a model designed by the office itself. The ideas of “active” and “passive” metadata were described by Jean-Pierre Kent and some of his colleagues.

29. The Semantic Web is the latest technological challenge to face the statistical offices. The desire to design services that know where to go to find relevant data is driving the work. Users won’t need to know the specific office or site where the data they want is stored. Combining data sets from different sources will be possible. But, this requires significant upgrades to data and metadata systems, such as using standards for interoperability and designing “computable” terminology systems (ontologies). This is the subject of the next section.

30. The other main focus of metadata systems within a statistical office is for driving internal processing. Interestingly, once the Internet was in use, NSOs began to look harder at internal metadata systems. In the few years before the advent of the Internet, local area and wide area networks were put into use in various organizations, and the Internet just expanded the trend. But, the new inter-connectivity made it possible to tie systems together in ways that had not been tried before. This meant metadata was needed in ways that were not evident before.

31. Inter-connecting systems meant that issues of data and system interoperability, process automation, “active” metadata (paragraph 28), metadata standards, and the survey process got much more visibility. Throughout the latter half of the 1990s, METIS saw many developments in this area from the statistical offices. By the year 2000, new standards efforts such as the DDI started focusing on the survey life-cycle. In the last few years, Part C of the CMF, and the Generic Statistical Business Process Model in particular, is a much

stronger focus for METIS and the Steering Group. Part A of the CMF also reflects the need for a strong internal focus.

V. TERMINOLOGY

32. The importance of terminology was recognized early as one of the main outputs of METIS. Early work focused on the development of a terminology for metadata systems, edited by Dusan Prazenka then Daniel Gillman, (published by the UNECE secretariat in 1999) and models for handling classification systems (a form of a terminological system), which evolved into the Neuchâtel Classification model (published in 2000 and available in the CMF web site through http://www.unece.org/stats). The METIS terminology, in turn, became one of the starting documents for the Metadata Common Vocabulary (MCV), edited by Marco Pellegrino and Denis Ward, in the Statistical Data and Metadata eXchange (SDMX) effort (see http://www.sdmx.org) under the aegis of the SDMX sponsor agencies (BIS, ECB, Eurostat, IMF, OECD, UNSD, and World Bank).

33. In general, terminology management consists of defining and relating the concepts underlying terms that are used in an office and gathering together all the terms used for one concept. There are a wide variety of possible uses and applications of this technique, and statistical offices have long recognized the importance of this. Classifications, thesauri, ontologies, nomenclatures, glossaries, and dictionaries are all examples; and most if not all of these kinds of structures are in use in statistical offices.

34. Many statistical offices were organizing their classifications in locally defined classification servers and building systems to organize the terminology their surveys use. Around 1999 it was recognized that standards for handling classifications and terminology would increase the chances of sharing these resources with others (i.e., increase interoperability), and discussion of these techniques and specifications followed. In Europe especially, the need for multilingual terminological systems was deemed important, and thesauri implementing these needs were started using *ISO*[9] standards (*ISO* 2499 and *ISO* 5788). These efforts are still a main area of work in many offices around the world.

35. More recently, the needs for more formalized terminological systems have been identified with the Semantic Web effort. These new terminological systems are known as ontologies, a term borrowed from philosophy, and they are essentially sets of related concepts with a known way to navigate the system. Navigation, in the most formal situations, allows for logical inferences to be made. If it is possible to build such systems for official statistics, it may be possible to implement a Semantic Web of statistical agencies. Many offices around the world are engaged in this kind of work presently. The needs for formalizing terminology make the effort complicated, and formalism is another theme throughout the history of METIS, which we look at next.

VI. FORMALISM

36. By formalism we refer to mathematical like structures imposed onto some object. Computer scientists are eager to try to solve problems with this technique. It means that whatever object they have described, they can compute with the description. If it is possible to achieve this, then a powerful system will result. The main limitation is whether the formal structure adequately reflects the object being described.

37. The early work in METIS made strong efforts to formalize many aspects of statistical metadata, including statistical indicators (Josef Olenski and Ebbo Petrikovits), statistical classifications (Galina Kolomyetseva), statistical metainformation systems (Bo Sundgren), and a data architecture (Markuu Saijets). These efforts and others were a major part of the early meetings.

---

[9] *ISO* – International Organization for Standardization. This is in italics to differentiate it from the other use of ISO (international statistical office).

38. By the latter half of the 90s, the focus was more on the models for statistical metadata (CLASET, Corporate Metadata Repository, Neuchâtel classification, Nordic model, and many others). The focus was less on a formal description than on a practical model for building systems. There was some formality underneath the descriptions, but the additional claims a formal system needed were not claimed. Moreover, many statistical offices were building systems using their own models, often based on previous work. At the same time, XML started to become very influential.

39. In the 2000s, the same trend continued (MetaNet, Neuchâtel variables, *ISO/IEC* 11179, SDMX, DDI, and others). *ISO/IEC* 11179 and DDI were developed as standards, and SDMX became one. SDMX and DDI were both based on XML and complied with *ISO/IEC* 11179. Both standards and models contain some formalness in their specifications, but they aren't usually mathematically based. In addition, as mentioned in Section V, terminology standards started being applied around 2000, and these included *ISO* 2788 (monolingual thesaurus), *ISO* 5964 (multilingual thesaurus), RDF (resource description framework), *ISO/IEC* 13250 (topic maps), and others.

40. The CMF Part B (Concepts, Standards, and Models) and Part C (Statistical Life-Cycle and Generic Statistical Business Process Model) are further efforts to advance the work in this area. Again, the formality of this work is less than the mathematical precision required by some, but it has a practical value we will visit in the next section.

41. Now with efforts to advance the Semantic Web, new kinds of formalism are being applied. RDF, OWL (Web Ontology Language), and Topic Maps are called formal description techniques and are being applied in statistical offices to accomplish the goal of providing more automatic methods for selecting and using statistical data. Papers on this subject are being presented in this METIS meeting.

VII. IMPLEMENTATION

42. The real focus of METIS is to help statistical offices design, build, and use statistical metadata systems. Even though theory, formalism, concepts, standards, and models are part of the discussion, they are not the final product of the work. Tracing meeting reports back to the earliest METIS sessions, much of the time was devoted to describing systems that were built, being built, or being planned. Every METIS meeting going back to 1993 has about half of the presentations devoted to the practical experience of implementing metadata systems in statistical offices.

43. As some humorist once said "In theory, theory and practice are the same. In practice, they are not." Though this statement may not be completely transparent to those who are not fluent in English, it relies on a meta-level – two uses of the words theory and practice – and it means that no matter how much a system is planned, implementing almost certainly requires some modifications. The experience implementers derive from building their systems and conveyed to participants during the meetings has been the most consistent aspect of METIS over the years. The feedback given to those who are interested in developing the theory has been invaluable.

44. Part D of the CMF is titled Implementation, and the explicit focus on that aspect of the work is testament to its importance. Since the work on the CMF was started in 2004, a much stronger effort to include the experiences of implementers has been realized. The 2007 and 2009 interim meetings were much more devoted to actual systems than the full meetings have the time for.

45. At the same time, Part A of the CMF is titled Statistical Metadata in a Corporate Context: A guide for managers (Jana Meliskova and Graeme Oakley). No broad implementations at the statistical office are going to happen or succeed without the support of management. The explicit acknowledgement of this need is vital to the success of incorporating statistical metadata management and systems in the office.

46. Over time, the complexity of the implementations in statistical offices has increased. Whereas, as we discussed in section IV-B, early implementations were more focused on data sets or single surveys, the

corporate approach is much more common today. This migration over time is evidence that the lesson about not building cathedrals was learned. A slow approach is much more effective. In fact, it is evidence of the idea we are trying to convey in paragraph 43.

47. With the new effort to incorporate ideas from the Semantic Web, corporate implementations will get more complicated. Given the careful approach taken so far by statistical offices in developing their statistical metadata systems, even if these efforts don't succeed, the overall systems will not fail. Seeing that this is the case, we conclude that METIS has been a great success so far. Let us ensure the next 20 years be as successful.